Education and Information Technologies (2022) 27:12381–12397
https://doi.org/10.1007/s10639-022-11117-1

Application of logistic regression to predict the failure of students in subjects of a mathematics undergraduate course

Stella F. Costa[1] · Michael M. Diniz[1]

Received: 15 February 2022 / Accepted: 9 May 2022 / Published online: 30 May 2022

Abstract

The large rates of students' failure is a very frequent problem in undergraduate courses, being even more evident in exact sciences. Pointing out the reasons of such problem is a paramount research topic, though not an easy task. An alternative is to use Educational Data Mining techniques (EDM), which enables one to convert data from educational database into useful information, in order to understand and improve teaching and learning processes. In this way, the objective of this paper is to propose mathematical models based on EDM techniques to estimate the probability of a student in a mathematics degree course at IFSP (Federal Institute of São Paulo) to fail in exact sciences disciplines, and later on, indicate which aspects contribute significantly for the Students' failure rates in these branches. We present three logistic regression models that which were applied based on socioeconomic data and student performance over 4 years. For interpretation and evaluation of such models, odds ratio, ten-fold Cross Validation method and the metrics: accuracy, sensitivity, specificity and area under the ROC curve (AUC) were used. It was noted that through Cross Validation, the models achieved accuracy values accounting for over 70%, sensitivity over 70%, specificity over 60% and AUC over 0.75. Analyzing the predictive variables of these models, we identified that factors such as advantage age, rates of failure through the course and attendance in initial semesters can increase the probability of failure in exact science disciplines in the analyzed course.

Stella F. Costa and Michael M. Diniz contributed equally to this work.

✉ Stella F. Costa
steefaria@gmail.com

Michael M. Diniz
michael.diniz@ifsp.edu.com.br

[1] Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - IFSP, Rod. Pres. Dutra, São José dos Campos 12223-201, SP, Brazil

Keywords Educational data mining · Logistic regression · Student performance · Data analysis

1 Introduction

The high rates of students' failure is a frequent and relevant problem in higher education courses around the world. Many researches have already been conducted in order to identify the factors that influence a student's performance throughout a course (Asif et al., 2017). In general, the results indicate that student performance is related to family factors, financial difficulties and ability with the content of the disciplines (Abu Saa et al., 2019).

For instance, in Aina (2013) the authors study the influence of parental factors on student performance in Italian universities, the findings indicate a positive correlation between the parents' education level and the student's performance.

In DesJardins et al. (2002), the authors show that universities with financial assistance programs for students managed to reduce the number of dropouts and increase the number of students' approvals. The positive effect of assistance programs on student's performance occurs more expressively in the first year of the course.

According to Barbosa and Concordido (2009) and Kato et al. (2015), the proportion of failures is even more relevant in disciplines in the field of Exact Science. The deficit in math skills of higher education of first-year students is a common and recurrent problem, Parsons (2004) shows that, due to these problems, students of Harper Adams University College demonstrated difficulties and high failure rates in subjects with mathematical and statistics content.

Calculus disciplines usuallypresent the highest proportion of failures among disciplines of Exact Science in undergraduate courses (López-Díaz and Peña, 2022). For instance, according to Barufi (1999), between 1990 and 1995, the failure rates in disciplines of Calculus in the University of São Paulo - Brazil exceeded the average rate of 50%, which was much greater than what happened in other disciplines from the same period. In Rezende (2003), it is shown that between 1996 and 2000, the rates of failure in Calculus at Fluminense Federal University, ranges between 45% and 95%. Researches like Machado (2008) and Wu (2018) discuss some factors that might influence such high failure rate in this discipline, as the lack of a cognitive structure capable of absorbing the complexity of the content; the use of ineffective teaching methods; and the lack of mastery on the necessary prerequisites for their learning. Finally, Silva et al. (2016) points out that difficulties faced by students in disciplines such as Calculus are not easily overcome, and lead to failure, retention and future dropping out of the course.

In order to know and understand the reasons that lead a student not to perform well, it is essential for teachers and managers to promote actions aiming to mitigate student failure. Nowadays, we have resources to store and manipulate a significant amount of data related to students' social reality and their academic performance. Thus, making it viable and promising the use of data mining techniques to extract knowledge and information about students' needs.

The study of data mining techniques in educational contexts is an emerging research area known as Educational Data Mining (EDM) (Romero & Ventura, 2010; Namoun & Alshanqiti, 2020). The main objective of EDM is to convert data from educational systems into useful information that can have a significant impact on research and practice, such information can be used in order to understand and improve student's teaching-learning process.

In literature, it is possible to find several researches involving EDM, for example, in Kovacic (2010), using CHAID and CART (decision tree algorithms), the authors evaluate the influence of variables related to the study environment and sociodemographic conditions, on students' dropout from the Information Systems course at Open Polytechnic of New Zealand. In Goldfinch and Hughes (2007), a regression model is used to explain the performance of freshman students using data about their self-confidence and learning styles. In Bhardwaj and Pal (2011), a model based on decision tree algorithms is proposed (using C4.5, ID3 and CART) to forecast students' individual performance in their final exams, finally, in George et al. (1994), logistic regression is used to predict students' success in subjects of an engineering course. Other examples of EDM implementation can be found on Al-Radaideh et al. (2006), Garman et al. (2010), and Henning et al. (2015).

Pointing out the reasons that lead students to fail is a topic that is increasingly drawing the attention of researchers and teachers (Fernandes Filho, 2001; Lopes, 1999; Pereira, 2018). In this context, it is believed that the number of failures and, consequently, dropouts can be minimized if, from the beginning of a discipline, some preventive actions are taken. Some important questions must be answered to guide the definition of these preventive actions, such as:

- Which students have the highest probability of failing?
- Which are the main reasons that culminate in students' failing?

For this reason, in this paper we will study the application of mathematical modelling to predict the probability of a specific student to fail in Exact Science disciplines. The study will be based on data of students enrolled at the Federal Institute of São Paulo (IFSP) in a teaching math undergraduate course, headquartered in the city of São José dos Campos.

Specifically, we will propose three different models of logistic regression, the first is to be focused on first-yeardisciplines, the second disciplines offered between the third and eighth semester, and the last one for disciplines offered all along the course. The three models will be trained based on students' socioeconomic features and student performance data.

Finally, a comparison will be made between the three proposed models, and based on these models, the variables that most contribute to student failure in the exact sciences subjects of the course will be presented.

The sections of the article are organized as follows: in Section 2, we describe the case of study and show some exploratory information about the student's performance history, in Section 3 the methodology used to develop the models, and the main features of the dataset used, are explained. In Section 4, the results are

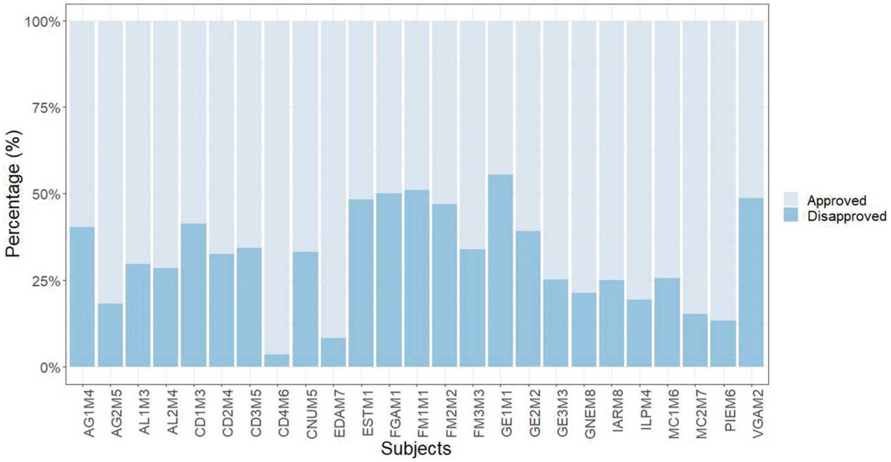


Fig. 1 Bar graph of failures and approval in exact sciences subjects between 2016 and 2019

presented and a comparison of the three proposed models is performed. Finally, in Section 5, we promote a discussion about the main results of the research.

2 Case description

The IFSP is a public institution that offers courses of secondary, professional and higher education level. The student selection for higher education courses is done through SISU (Unified Selection System) and managed by Educational Ministry (MEC) which uses the ENEM (High School National Exam) score as selection criteria.

The IFSP has 37 campuses distributed in the state of São Paulo. One of these is headquartered in the city of São José dos Campos, which we will denote along the text by IFSP-SJC. Among other courses, the IFSP-SJC offers a teaching math undergraduate course, which has an expected duration of 4 years and offers 40 new vacancies every beginning of the year.

A serious problem of this course is the high number of failures in disciplines of exact sciences.The graph in the Fig. 1 shows the percentage of failures in each discipline of exact science in the course between 2016 and 2019. In more than 80% of subjects, the number of failures is greater than 20%.

Subjects of the first semester as FM1M1 (Fundamentals of Mathematics 1), GE1M1 (Geometry 1), FGAM1 (Fundamentals of Analytical Geometry 1) and ESTM1 (Basic Statistics) had approximately 50% of their students failing, which shows that the problem is still more accentuated in first semester disciplines.

It can also be seen from Fig. 2 that the percentage of failures in the first two semesters is significantly higher than in the remaining semesters.

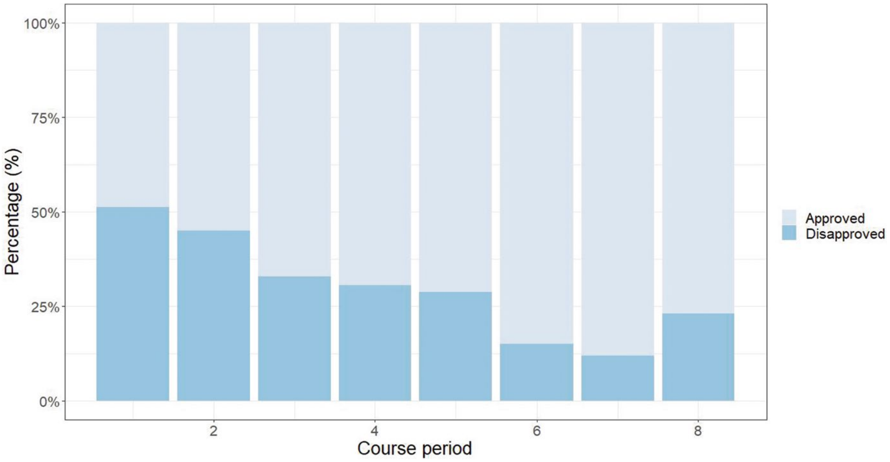


Fig. 2 Bar graph of failures and approvals in exact sciences subjects according to the period. The number of failures among students of first semester exceed 50%. In the other semesters, the percentage of failures decreases accounting for 45%, 36.9%, 31.57%, 28.71%, 15.05%, 12% e 23.08%, respectively

The main objective of this research is to enable a decrease in the number of failures by identifying the reasons that cause students' flop and identify those who are most vulnerable to be reproved.

3 Methodologies

3.1 Models

According to James et al. (2013), considering the Bernoulli random variable Y ($Y=0$ or $Y=1$), a logistic regression model is given by the logistic function as follows

$$p(Y=1)=\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_n x_n}}, \tag{1}$$

where x_1, x_2,..., x_n are the predictor variables of the model, β_0, β_1,..., β_n are coefficients to be determined, n is the quantity of predictor variables and $p(Y=1)$ it is the probability of $Y=1$.

The choice of the logistic regression models is justified by the type of available data and by the nature of expected answer of the models, i.e., we seek to analyze the relationship between a binary dependent variable Y (0 - approval or 1 - failure) and a set of numerical and categorical independent variables (Hosmer et al., 2000).

Our models have 14 predictor variables that are listed in Table 1 and will be detailed in the Section 3.3.

In this article we developed three logistic regression models as follows.

Table 1 List of considered variables

Variable name	Variable type	Response category
Sex	Categorical	0-Female, 1-Male
Age	Numerical	{17,18,19,...}
ENEM grade	Numerical	{100,101,...,999,1000}
High School Type (HST)	Categorical	0-Private school, 1-Public school
Course Period (CP)	Numerical[a]	{1,2,3,...,7,8}
First Year Subject (FYS)	Categorical	0-No, 1-Yes
Failures in the subject due to absence (FSA)	Numerical	{0,1,2,...}
Failures in the subject due to bad grade (FSG)	Numerical	{0,1,2,...}
Failures in the subject (FS)	Numerical	{0,1,2,...}
Failures in course due to absence (FCA)	Numerical	{0,1,2,...}
Failures in course due to bad grade (FCG)	Numerical	{0,1,2,...}
Failures in the course (FC)	Numerical	{0,1,2,...}
Student achievement index (SAI)	Numerical	[0, 10]
Enrollment time in the course - (ET)[b]	Numerical	$[0, +\infty[$

[a] For Model 1, this variable was considered categorical and the response categories are: 1 - 1st period, 0 - 2nd period

[b] at the beginning of the subject (in years)

- Model 1: only considers students enrolled in the first year;
- Model 2: only considers students enrolled in second, third or fourth year subjects;
- Model 3: considers all students enrolled in the course.

In general, first-year students form a more heterogeneous group than students in the remaining years of the course, thus, it is expected that the reasons that make a student fail in first year disciplines are not entirely equal to those that make students fail in remaining years. Therefore, we propose a specific model for first-year students (Model 1) and another for students of the remaining years (Model 2). The Model 3 will help to measure how effective a single model is in predicting student failure, considering all enrolled students.

The logistic model output is the probability of a student fail in a specific discipline, to get a binary answer, we set a cut point p_0, so that, for $p(X) \geq p_0$ the models indicate student failure, on the other hand, for $p(X) < p_0$ the models indicate student approval.

3.2 Dataset

The data used in this research was extracted from SUAP (Unified Public Administration System) and academic records office. These records are about students enrolled in the course between 2016 and 2019. The data brings information related to admission, socioeconomic characteristics and student performance in disciplines of exact science.

On original database, the final status of each student in a discipline could receive the following ratings:

- **Disapproved**: the student reproved by grades or absence;
- **Approved**: the student met all approval criteria;
- **Canceled**: the student whose enrollment in the course was canceled throughout the semester;
- **Took time off**: the student took time off from discipline;
- **Excused**: the student was excused from taking the discipline by excused equivalence policy.

We removed the data referring to canceled, excused and students that took time off, therefore, we consider as disapproved only students reproved by grades or absence.

The final dataset has 1878 observations, of which 1061 observations were used for Model 1, 817 observations for Model 2 and all for Model 3.

3.3 Pre-processing

The original dataset has 17 attributes. However, due to the high percentage of missing data the attributes "per capita income", "gross family income" and "type of transport" were discarded. The 14 remaining attributes are described in the Table 1.

The variable HST refers to the administrative dependence of the school, i.e., if the High School attended by the student is public or private. The variable SAI is the average of grades obtained by the student in the subjects in which he was approved, and the variable ET is the time, in years, of enrollment. The variables Sex, HST and FYS are categorical (Table 1), and the chosen reference categories were "Female", "Private school" and "No", respectively.

In order to balance the dependent variable, the ROSE technique was applied (Lunardon et al., 2014) in Models 2 and 3. Before ROSE, the dataset for Model 2 had 71.4% of approved students and 28.6% of reproved. In the dataset for Model 3 the proportion consisted of 59.8% "Approved" against 40.2% "Reproved". After ROSE, the proportion comprised 49% "Approved" and 51% "Reproved" for dataset 2, and 51.4% "Approved" in contrast to 48.6% "Reproved" for dataset 3.

3.4 Model development and evaluation

We split the dataset between 30% for testing and 70% for training and validation. For training and validation, the method K-folds cross validation was performed with $K = 10$ (Cunha, 2019; Kohavi, 1995). The models were analyzed and evaluated based on the criteria listed in the following sub-section.

3.4.1 Odds ratio

Let E be an arbitrary event and $p(E)$ their respective probability, the Odds of E are given by

$$odds(E) = \frac{p(E)}{1 - p(E)}.$$

In the logistic model, the odds of $E = S$, where S is the sucess event, depends on the vector of independent variables $X = (x_1, x_2,..., x_p)$, therefore, the odds of $E = S$ given X ($odds(S/X)$) are:

$$odds(S/X) = \frac{p(S/X)}{1 - p(S/X)} = \frac{\frac{e^{\beta_0+\sum\beta_i x_i}}{1+e^{\beta_0+\sum\beta_i x_i}}}{1 - \frac{e^{\beta_0+\sum\beta_i x_i}}{1+e^{\beta_0+\sum\beta_i x_i}}} = \frac{\frac{1}{1+e^{-(\beta_0+\sum\beta_i x_i)}}}{\frac{e^{-(\beta_0+\sum\beta_i x_i)}}{1+e^{-(\beta_0+\sum\beta_i x_i)}}} = e^{\beta_0+\sum\beta_i x_i},$$

where $\sum \beta_i x_i$ is a suppressed notation for

$$\sum_{i=1}^{p} \beta_i x_i.$$

Considering two vectors of independent variables X_0 and X_1, the Odds Ratio (OR) is given by

$$OR(X_1, X_0) = \frac{odds(S/X_1)}{odds(S/X_0)} = \frac{e^{\beta_0+\sum\beta_i x_{1i}}}{e^{\beta_0+\sum\beta_i x_{0i}}} = e^{\sum\beta_i(x_{1i}-x_{0i})}.$$

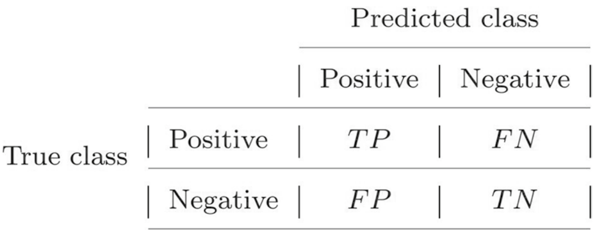
The OR is the ratio between the chances of a given outcome occurring, considering two possible sets of independent variables X_0 and X_1. When the difference between X_0 and X_1 occurs in a unique variable, the OR indicates how much and how this variable influences the probability of occurrence of the outcome.

In the scope of this paper, the OR indicates how much each predictor variable influences the probability of students' failure.

3.4.2 Metrics of performance

The following diagram is known as "Confusion Matrix", and depicts a summary of a model performance (Fig. 3).

Fig. 3 Confusion Matrix diagram: TP - True Positive, FN - False Negative, FP - False Positive and TN - True Negative



		Predicted class	
		Positive	Negative
True class	Positive	TP	FN
	Negative	FP	TN

From Confusion Matrix, some performance evaluation metrics are defined. The "sensitivity" is the ratio between true positives (*TP*) and total of positives (*TP* + *FN*), i.e., the sensitivity measures the model's ability to classify an input as positive given that it really is positive.

$$Sensitivity = \frac{TP}{TP + FN}$$

The "specificity" is the ratio between true negatives (*TN*) and total of negatives (*TN* + *FP*), i.e., the specificity measures the model's ability to classify an input as negative given that it really is negative.

$$Specificity = \frac{TN}{TN + FP}$$

The "accuracy" of a model, is the proportion of correct classifications.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Finally, one way to assess the performance of a logistic regression model is through the Receiver Operating Characteristic curve (ROC) (Bradley, 1997; Spackman, 1989).

The ROC curve is obtained by representing the true positive rate (sensitivity) on x-axis versus the false positive rate (1-specificity) on y-axis for p_0 varying between 0 and 1.

The area under the ROC curve (AUC) ranges between 0 and 1, this metric indicates the model's capacity to correctly differentiate between success and failure cases (Hosmer et al., 2000). Values close to 1 indicate that the model has good performance (Table 2).

3.4.3 Method to select variable

To check whether a given independent variable has a statistically significant relationship with the dependent variable, the Wald Test is performed (Cabral, 2013). Basically, the Wald Test verifies if each coefficient of the logistic model is equal to zero or not. The hypothesis tested are:

Table 2 Levels of discrimination power of the model as a function of the AUC (Hosmer et al., 2000)

AUC $=0.5$	This suggests no discrimination.
$0.7 \leq$ AUC < 0.8	This is considered acceptable discrimination.
$0.8 \leq$ AUC < 0.9	This is considered excellent discrimination.
AUC ≥ 0.9	This is considered outstanding discrimination.

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}, i = 0, 1, ..., n$$

where n is the number of independent variables of the model.

Therefore, when the p-value associated with a variable is less than the significance level α, we reject the null hypothesis and conclude that there is, in fact, an association between the variable and the probability of failure, i.e., the independent variable is considered significant to the model.

3.5 Model development flowchart

In summary, the study presented in this article was performed according to the steps of the flowchart in Fig. 4.

The steps of the Model development flowchart are explained as follows:

1. **Problem analysis**: The problem that should be solved by the model is defined and a general analysis about relevance, implications and possible difficulties is stated. In this step, any type of study on the problem is welcome.
2. **Determination of possible predictor variables**: All variables that can hypothetically influence the outcome of the model are listed. In this step, the researcher should not consider the accessibility or availability of data for each variable.
3. **ETL process**: Extraction-Transformation-Loading (ETL) processes (Vassiliadis et al., 2002) is the process of transforming raw data into suitable dataset for training a model. In this step, data extraction, balancing, aggregation of different database, completion of missing data, transformations and etc are performed. Commonly, in this step, it is noted the impossibility of using some variables listed in Step 2 of the flowchart.
4. **Dataset split**: The dataset is partitioned into training data and test data. Training data are used throughout the model training process, while test data are used only to evaluate the final model obtained.
5. **Model determination**: In this step, not only is the best model type chosen, but also its hyperparameters. The method commonly used in this step is cross validation (Cunha, 2019; Kohavi, 1995). Here, only training data are used.

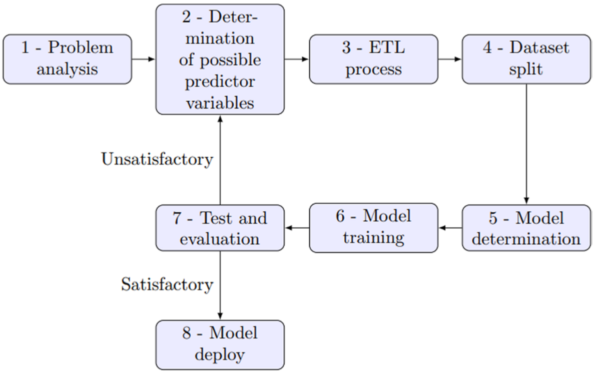


Fig. 4 Model development flowchart

6. **Model Training**: The model is trained with all training data. In this step, a final model is obtained.
7. **Test and evaluation**: The trained model is tested in the test data. If the performance is unsatisfactory, a revision in all the process from Step 2 is necessary, or it is concluded the impossibility of solving the problem established in Step 1.
8. **Model deploy**: If the tests of Step 8 have been satisfactory, the model is used in real situations and is available to support decision-making in the established context.

The flowchart of Fig. 4 enables the reproduction of the study presented in this article to other environments or educational spaces with different characteristics. Not limiting the methodology to develop student performance forecast models, but enabling its replication to other themes in the educational context.

4 Results

The results of the construction and evaluation steps of the three proposed models will be presented.

4.1 Model parameters and variable selection

According to the variable selection method presented in Section 3.4.3, the predictor variables *Age, ENEM grade, CP, FS* and *FC* significantly influence the probability of a student failing in first-year subjects. Considering the Eq. 1, the selected variables (x_i) and their respective parameters (β_i) are presented on Table 3.

Checking the OR (Table 4), we can observe that a student who is taking first semester subjects is almost 3 times more likely to fail than a student who is in the second semester. Futhermore, the older a student is or the more their failures already obtained in the course are, the larger is probability of failing. Finally, the higher the ENEM grade or the number of failures already obtained in the subject, the lower the chances of failure.

As for Model 2, the predictor variables *Age, FCA, FCG, SAI* and *ET* significantly influence the probability of a student failing the second, third, and fourth-year subjects. Considering the Eq. 1, the selected variables (x_i) and their respective parameters (β_i) are presented on Table 5.

Table 3 Predictor variables (x_i) pointed as significant for Model 1 and their respective parameters (β_i)

Variable (x_i)	Estimate (β_i)	Std. Error	z value	p-value	2.5%	97.5%
(Intercept)	3.6774	1.0379	3.543	3.95e-04	1.6698	5.7457
Age	0.0306	0.0097	3.167	1.54e-03	0.0119	0.0500
ENEM grade	–0.0088	0.0016	–5.663	1.49e-08	–0.0120	–0.0058
CP	1.0030	0.2059	4.872	1.10e-06	0.6060	1.4143
FS	–3.6583	0.6569	–5.569	2.57e-08	–5.0152	–2.4315
FC	0.7206	0.1022	7.054	1.74e-12	0.5301	0.9322

Table 4 The Odds Ration (OR) for each selected variable (x_i) and their respective 95% confidence interval

Variable (x_i)	OR	2.5%	97.5%
(Intercept)	39.544	5.311	312.854
Age	1.031	1.012	1.051
ENEM grade	0.991	0.988	0.994
CP	2.726	1.833	4.114
FS	0.026	0.007	0.088
FC	2.056	1.699	2.540

Table 5 Predictor variables (x_i) pointed as significant for Model 2 and their respective parameters (β_i)

Variable (x_i)	Estimate (β_i)	Std. Error	z value	p-value	2.5%	97.5%
(Intercept)	1.1951	0.5324	2.245	2.48e-02	0.1850	2.2761
Age	0.0459	0.0100	4.567	4.94e-06	0.0266	0.0661
FCA	0.3560	0.0583	6.101	1.06e-09	0.2454	0.4746
FCG	0.2874	0.0421	6.825	8.80e-12	0.2082	0.3736
SAI	–0.3463	0.0643	–5.383	7.32e-08	–0.4786	–0.2261
ET	–0.4595	0.1231	–3.733	1.89e-04	–0.7044	–0.2208

Thus, when analyzing the OR (Table 6) we conclude that the chance of a student failing increases according to the how old they are, analogously, the more disapprovals already obtained in the course, the greater are the chances of a new failure. On the other hand, when we have an upward trend in the IRA or in the time enrolled in the course, there is a reduction in the chance of failure.

As for Model 3, the predictor variables *Age, ENEM grade, CP, FCA, FCG, SAI* and *ET* significantly influence the probability of a student failing any subject in the course. Considering the Eq. 1, the selected variables (x_i) and their respective parameters (β_i) are presented on Table 7

Based on the OR (Table 8), we conclude that the older a student is or the larger is their number of failures already obtained in the course, the bigger is the probability of failure. As for the ENEM grade, the higher the student's grade, the lower are the chances of failure, as well as the increase in IRA and the time of enrollment in the course.

4.2 Model performance

In Table 9, the models performance is verified by applying the Cross-Validation method, which produced accuracy higher than 70%, sensitivity above 70%, specificity over 60%, and AUC greater than 0.75, indicating that Model 1 and Model 3 have acceptable discrimination, and Model 2 excellent discrimination.

To infer the behavior of each model in new data (not used for training and testing), test samples were utilized. The models performance in the test sample is shown in Table 10.

Table 6 The Odds Ration (OR) for each selected variable (x_i) and their respective 95% confidence interval

Variable (x_i)	OR	2.5%	97.5%
(Intercept)	3.304	1.203	9.739
Age	1.047	1.027	1.068
FCA	1.428	1.278	1.607
FCG	1.333	1.231	1.453
SAI	0.707	0.620	0.798
ET	0.632	0.494	0.802

Table 7 Predictor variables (x_i) pointed as significant for Model 3 and their respective parameters (β_i)

Variable (x_i)	Estimate (β_i)	Std. Error	z value	p-value	2.5%	97.5%
(Intercept)	4.2691	0.6704	6.368	1.91e-10	2.9694	5.5994
Age	0.0357	0.0064	5.572	2.51e-08	0.0233	0.0485
ENEM grade	–0.0055	0.0010	–5.682	1.33e-08	–0.0074	–0.0036
CP	–0.2034	0.0432	–4.712	2.46e-06	–0.2887	–0.1193
FCA	0.1732	0.0430	4.032	5.52e-05	0.0907	0.2592
FCG	0.1118	0.0300	3.727	1.94e-04	0.0539	0.1717
SAI	–0.1795	0.0289	–6.206	5.44e-10	–0.2375	–0.1240
ET	–0.2652	0.0839	–3.162	1.57e-03	–0.4307	–0.1015

Table 8 The Odds Ration (OR) for each selected variable (x_i) and their respective 95% confidence interval

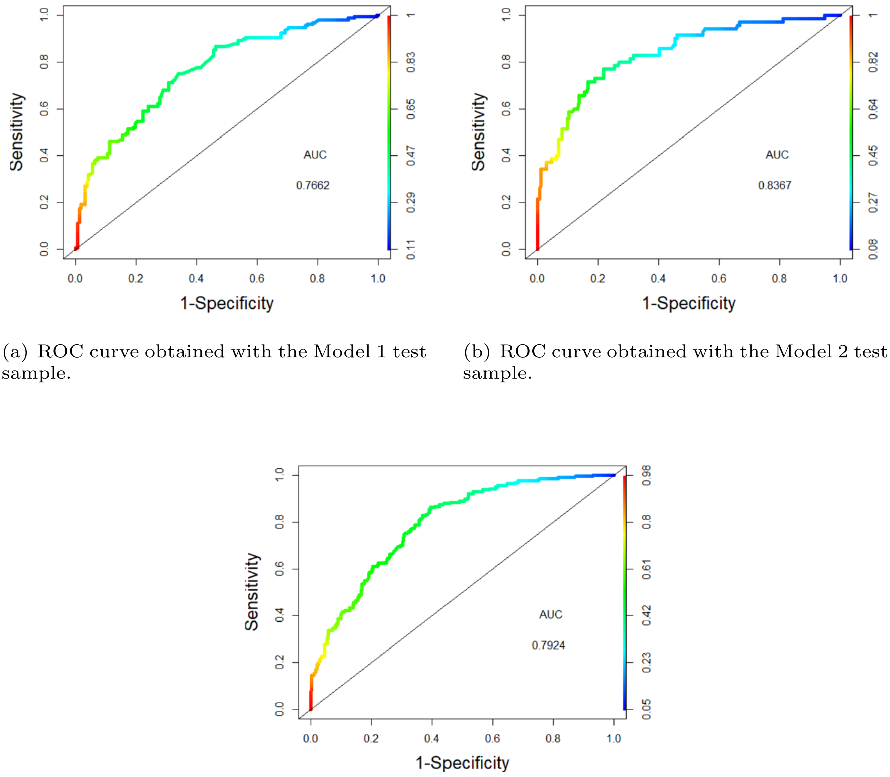
Variable (x_i)	OR	2.5%	97.5%
(Intercept)	71.461	19.480	270.266
Age	1.036	1.024	1.050
ENEM grade	0.995	0.993	0.996
CP	0.816	0.749	0.888
FCA	1.189	1.095	1.296
FCG	1.118	1.055	1.187
SAI	0.836	0.789	0.883
ET	0.767	0.650	0.903

Table 9 Model performance in the training set using ten-folds cross validation

Models	Accuracy	Sensitivity	Specificity	AUC
Model 1	0.7035	0.7798	0.6199	0.7618
Model 2	0.7815	0.8448	0.7130	0.8277
Model 3	0.7050	0.7196	0.6953	0.7691

Table 10 Models performance on test set

Models	Accuracy	Sensitivity	Specificity	AUC
Model 1	0.673	0.6218	0.7222	0.7662
Model 2	0.791	0.6857	0.8333	0.8367
Model 3	0.7069	0.6637	0.7359	0.7924



(a) ROC curve obtained with the Model 1 test sample.

(b) ROC curve obtained with the Model 2 test sample.

(c) ROC curve obtained with the Model 3 test sample.

Fig. 5 ROC curves obtained from the test set. Analogous to the result obtained from Cross Validation, curves (**a**) and (**b**) indicate acceptable discrimination of the models, and curve (**c**) indicates excellent discrimination

When analyzing the results obtained from applying the models to the test set, it is verified that the models achieved values for accuracy above 65%, the proportion of true positives consisting of over 60%, the proportion of true negatives comprising higher than 70%, and AUC higher than 0.75. We can observe that the performance of Model 2 was superior to the others in all metrics analyzed (Fig. 5).

The results presented in Table 9 are similar to the results shown in Table 10, which indicate reliability in the predictions made by such models.

5 Discussions and conclusions

According to the Tables 3, 5 and 7, the independent variables age, ENEM grade, CP, FS, FC, FCA, FCG, SAI and ET (see Table 1) significantly influence the probability of a student failing the subjects assessed. Moreover, the aspects that most imply in the increase of the probability of students' failure are: old age, high number of failures in the course (regardless of the reason) and attending the initial semesters.

As expected, while comparing Tables 3 and 5, we see that the variables used to predict student failures in the first year (Model 1) are completely different from those used to predict performance of the remaining-years students (Model 2). In relation to Model 3, the set of predictor variables is practically the union between the predictor variables of Model 1 and 2, only not including the variables "Failures in the subject (FS)" and "Failures in the course (FC)".

Based on the accuracy and AUCs obtained using Cross Validation and on the test samples (Tables 9 and 10), we concluded that Model 2 was the one with the best performance. On the other hand, Model 3 performed slightly better than Model 1. Therefore, confirming that it is more difficult to predict the failure of students from the first year than for those of the second year onwards. Even so, Model 1 presents suitable performance results and can contribute to reduce the number of failures in the most critical moment of the course, which is the first year.

First-year students form a very heterogeneous group in terms of performance and affinity with disciplines of exact area, on the other hand, students from the second year onwards, in general, have already adapted to the course routine, developed efficient study methodologies and aim to effectively complete the course, which makes the group more homogeneous and justifies the difference in performance between the models.

Observing Table 9, the models achieved an average sensitivity between 71% and 84%, this indicates that, in average between 71% and 84% of the students who will fail in a given subject are likely to be identified by the models as possible case of failure. Which means that most students who would fail (without intervention) will have the opportunity to receive adequate support throughout the course in order to avoid such. This metric performed more poorly in the test sample (Table 10), however, it still remained at levels that attest to its relevance.

Still considering the Table 9, it is possible to verify that the models reached an average specificity between 61% and 71%, which shows that, in average between 61% and 71% of students who would not fail are correctly classified. In the test sample, this metric reached values between 72% and 83%, which reinforces the efficiency of the proposed models.

With the results of this paper, it is noted that at the beginning of a discipline, students who are more likely to fail are identified, and thus, with the objective of minimizing the number of failures, some intervention measures can be taken, such as: raising awareness to the teaching and learning process of certain students; applying methodologies that help older students and using motivational strategies for those who fail in the first year

Finally, we believe that the performance of these models would improve if socioeconomic variables such as "per capita income", "gross family income" and "type of transport" had not had so much missing data and therefore could have been incorporated into the model.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Stella F. Costa and Michael M. Diniz. The first draft of the manuscript was written by Stella F. Costa and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This study was funded by IFSP's institutional program, PIBIFSP, process number 23305.016744.2019-35.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Ethics approval According to the Ethics Committee in Research with Human Beings of the Federal Institute of São Paulo (CEP-IFSP),[1] all projects which must be submitted to the Committee are defined in Article 1 of resolution 510/2016.[2] According to the aforementioned document (article 1, paragraph V and VII), it was not necessary to register and evaluate the activity proposed in this work, since it is a database research, whose information is aggregated, without the possibility of individual identification.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest/Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technology, Knowledge and Learning, 24*(4), 567–598.

Aina, C. (2013). Parental background and university dropout in Italy Parental background and university dropout in italy. *Higher Education, 65*(4), 437–456.

Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006). Mining student data using decision trees. In *International Arab conference on information technology (ACIT'2006)*. Jordan: Yarmouk University.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113,* 177–194.

Barbosa, A. C. D. C., & Concordido, C. F. R. (2009). Ensino colaborativo em Ciências Exatas. Ensino Saúde e Ambiente, 2(3).

Barufi, M. C. B. (1999). *A construção/negociação de significados no curso universitário inicial de Cálculo Diferencial e Integral*. São Paulo: FE–USP.

Bhardwaj, B. K., & Pal, S. (2011). Data mining: A prediction for performance improvement using classification (IJCSIS). *The International Journal of Computer Science and Information Security, 9*(4), 136–140.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 145–1159.

Cabral, C. I. S. (2013). Aplicação do modelo de regressão logística num estudo de mercado. School Tese de Mestrado. Universidade de Lisboa.

Available in: https://ifsp.edu.br/acoes-e-programas/106-reitoria/conselhos-e-nucleos/858-comite-de-etica-em-pesquisa-com-seres-humanos-cep

Available in: https://www.ifsp.edu.br/images/reitoria/Comites/RESOLUO-510-de07abril_2016---CONEP.pdf

Cunha, J. P. Z. (2019). Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos. School Tese de Mestrado. São Paulo Instituto de Matemática e Estatística da Universidade de São Paulo (IME - USP).

DesJardins, S. L., Ahlburg, D. A., & McCall, B. P. (2002). A temporal investigation of factors related to timely degree completion. *The Journal of Higher Education, 73*(5), 555–581.

Fernandes Filho, O. P. (2001). O desenvolvimento cognitivo e a reprovação no curso de engenharia. In: XXIX Congresso Brasileiro de Ensino de Engenharia, pp 15–22. Porto Alegre.

Garman, G., et al. (2010). A logistic approach to predicting student success in online database courses. *American Journal of Business Education (AJBE), 3*(12), 1–6.

George, G., Moore, E., & Patey, M. (1994). A simple model for predicting success in an engineering programme. *International Journal of Engineering Education, 10,* 268–268.

Goldfinch, J., & Hughes, M. (2007). Skills, learning styles and success of first-year undergraduates. *Active Learning in Higher Education, 8*(3), 259–273.

Henning, E., Moro, G., Pacheco, P. S., & Konrath, A. C. (2015). Fatores determinantes para o sucesso na disciplina de cálculo diferencial e integral aplicando a regressão logística. *Revista de Ensino de Ciências e Engenharia, 6*(1), 122–141.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). *Applied logistic regression*. New York: Wiley.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r*. New York: Springer.

Kato, L. A., Gerônimo, J. R., Cardoso, V. C., Zanella, M. S., Niro, K. L., & de Souza, J. T. G. (2015). Performance of first-year undergraduate students attending exact sciences courses in problems of the additive conceptual field. *Acta Scientiarum Education, 37*(4), 383–390.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, (Vol. 14 pp. 1137–1145).

Kovacic, Z. (2010). Early prediction of student success: Mining students enrolment data. Informing Science & IT Education Conference (InSITE).

Lopes, A. (1999). Algumas reflexões sobre a questão do alto índice de reprovação nos cursos de Cálculo da UFRGS. *Sociedade Brasileira de Matemática Rio de Janeiro, 26*(/27), 123–146.

López-Díaz, M. T., & Peña, M. (2022). Improving calculus curriculum in engineering degrees: Implementation of technological applications. *Mathematics, 10*(3), 341.

Lunardon, N., Menardi, G., & Torelli, N. (2014). Rose: A package for binary imbalanced learning. R Journal, 6(1).

Machado, S. (2008). *Teoria das situações didáticas*. EDUC (Série Trilhas): São Paulo.

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences, 11*(1), 237.

Parsons, S. J. (2004). *Overcoming poor failure rates in mathematics for engineering students: A support perspective*. Newport: Harper Adams University College.

Pereira, M. V. C. (2018). *Análise sobre os índices de reprovação nos cursos de Cálculo I da UFERSA Trabalho de Conclusão de Curso*. Rio Grande do Norte: Universidade Federal Rural do Semi-Árido (UFERSA).

Rezende, W. M. (2003). *O ensino de Cálculo: Dificuldades de natureza epistemológica Tese de Doutorado*. Universidade de São Paulo (USP): São Paulo.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601–618.

Silva, A. C., et al. (2016). Análise dos índices de reprovação nas disciplinas de Cálculo I e AVGA do curso de Engenharia Elétrica do Instituto Federal da Bahia de Vitória da Conquista. XIV International Conference on Engineering and Technology Education.

Spackman, K. A. (1989). Signal detection theory: Valuable tools for evaluating inductive learning. In: Proceedings of the sixth international workshop on machine learning, pp 160–163.

Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). Conceptual modeling for ETL processes. In: Proceedings of the 5th ACM international workshop on data warehousing and OLAP, pp 14–21.

Wu, X. (2018). Persistence and characteristics of calculus I students in STEM disciplines. West Virginia University.